FEBS Letters 581 (2007) 826–830

VPMCD: Variable interaction modeling approach for class discrimination in biological systems

Rao Raghuraj, Samavedham Lakshminarayanan*

Department of Chemical and Biomolecular Engineering, 4 Engineering Drive 4, National University of Singapore, Singapore 117576

Received 11 September 2006; revised 19 January 2007; accepted 25 January 2007

Available online 2 February 2007

Edited by Robert B. Russell

Abstract **Data classification algorithms applied for class prediction in computational biology literature are data specific and have shown varying degrees of performance. Different classes cannot be distinguished solely based on interclass distances or decision boundaries. We propose that inter-relations among the features be exploited for separating observations into specific classes. A new variable predictive model based class discrimination (VPMCD) method is described here. Three well established and proven data sets of varying statistical and biological significance are utilized as benchmark. The performance of the new method is compared with advanced classification algorithms. The new method performs better during different tests and shows higher stability and robustness. The VPMCD is observed to be a potentially strong classification approach and can be effectively extended to other data mining applications involving biological systems.**

Keywords: Data classification; Variable predictive models; Discriminant analysis; Computational biology; Machine learning; Multivariate statistics

1. Introduction

Multivariate data classification into different known characteristic groups is a significant problem with far reaching outcomes in many fields of science and engineering. Taxonomical studies, analysis of expression profiles, biomarkers identification, protein structure and function prediction and clinical diagnosis are just a few of the research areas of computational biology that benefit largely from the application of various supervised classification algorithms. Many classifier functions have been tried in the literature with varying degrees of success for different classification problems especially for bioinformatics applications [1,2]. The distance based similarity search methods [3], class separating decision boundary seeking methods; linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) [2,3], support vector machines (SVM) [4], information content based decision tree methods; classification and regression trees (CART) [5] and black box model based artificial neural networks (ANN) have all been tried [2,3]. A good review of significance and limitations of many classification algorithms used in literature can be found in [1–3,6].

Recent algorithms like SVM and neural networks provide excellent learning capabilities and have shown high self-consistency for a wide range of problems. Neural network is a slower algorithm as the computational load depends on number of classes, variables and data size. The kernel based SVM method is almost independent of feature space and can effectively handle low sample sizes. This ability has established the superiority of SVM method especially for large scale bioinformatics applications. Nevertheless, SVM employs rigorous optimization algorithms, whose performance can be sensitive to parameters defined in its kernel function. Moreover, extension of the basic binary SVM classifier to multi-class problems is not well established; it is also computationally more taxing. The performance of these existing methods is affected during more challenging applications while testing the classifier with data samples for which they were not trained. The existing discrimination methods also do not capitalize on the association between the predictor features which can bring distinct dissimilarities between classes. Such variable interactions characterizing the structure can be mathematically established and the distinct relations can be used as discriminating models. The new variable predictive model class discrimination (VPMCD) technique proposed here exploits this new concept for class discrimination. Such a classification approach is of higher significance to biological systems which are known to show interactions among components used to characterize the system.

2. Materials and methods

2.1. Feature association modeling using variable predictive models

Different types of system behavior are always quantified using measurable features and interactions among them. Correlation based methods can define such associations between continuous predictor attributes and have been widely employed in literature for many data mining problems [1,7]. However, this qualitative analysis cannot distinguish linear/nonlinear, direct/indirect relationships between variables. The variable associations require richer quantitative representations and mathematical insights for characterizing certain definitive system behavior. Such deterministic relations (termed here as variable predictive models (VPM)) can be suitably developed and validated from the observations made on the system. Consider a system N with measur-

*Corresponding author. Fax: +65 6779 1936.
E-mail address: chels@nus.edu.sg (S. Lakshminarayanan).

Abbreviations: VPM, variable predictive model; VPMCD, VPM based class discrimination; ULDA, uncorrelated linear discriminant analysis; SVM, support vector machines; CART, classification and regression trees; LOOCV, leave one out cross validation; nFCV, n fold cross validation; PSL, protein sub-cellular location; FC, Fisher criteria

0014-5793/$32.00
doi:10.1016/j.febslet.2007.01.052

able continuous attributes $[X_1, X_2, \ldots, X_p]$. A VPM_i defined for any continuous variable X_i in system N is a mathematical equation (with linear/nonlinear, univariate/multivariate structure) modeled using sample measurements of other attributes in N. The model VPM_i can potentially define variable X_i as a function of best set of other variables of the same system $(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_p)$. The basic assumption here is that certain measured variables (X_i and X_j; $j \neq i$), characterizing the system, are dependent. The designed models are validated using goodness tests based on the prediction errors. The model with highest degree of fitness during validation is selected as VPM_i for variable X_i. Each variable X_i in the system is thus modeled and the collection of these VPM_i (number equal to the number of variables in the system) is used as characteristic model representing the inter-variable associations.

Though the model structures for VPM depend largely on the nature of the system and variables selected to define it, simple polynomial models can potentially capture majority of variations and associations between variables. Four basic model types (Linear (L), Linear + Interaction (LI), Quadratic + Interaction (QI) and pure Quadratic (Q) model types) are adopted here to clarify the VPM concept and its further use for discriminant analysis. The definition and details of these four model types are shown in Table 1. The number of other variables used for prediction in VPM_i is referred as predictor order (r). Both univariate ($r = 1$) and multivariate ($r > 1$) models are used for the above four types of VPMs. Once the model type and predictor order (r) are chosen, the parameters 'B' are estimated using the training data set. A linear regression problem is formulated as $Y = D \cdot B$ to determine the set of 'b' values (in vector B) using ordinary least squares [8]. Here Y is the variable to be predicted ($n \times 1$; column corresponding to X_i, the variable being predicted in the training set), B is the model parameter vector ($q \times 1$) and D is the design matrix ($n \times q$) with the polynomial values of predictor variable set (X_j; $j \neq i$). The description of design matrix and subsequent number of parameters (q) are outlined for each type of model in Table 1. It should be noted that each variable X_i ($i = 1, 2, \ldots, p$) in the given system N takes a role of predicted variable (Y) and for each X_i some of the remaining variables are used as predicting variable set X_j ($j = 1, 2, \ldots, r$; $j \neq i$). The regression coefficients for linear terms (b_j and b_{1k}), quadratic terms (b_{2k}) and interaction terms (b_{kl}) used in the second column of Table 1 are obtained from the least squares formulation using the training data set N. These (standardized) regression coefficients signify the contribution of each predictor variable on the corresponding predicted variable X_i.

2.2. VPM based class discrimination (VPMCD) as applied to data classification

If a system exhibits different classes of behaviors, it can be hypothesized that the structure of associations between the same set of variables will also be different in each of the classes. This provides the basis for extending the VPM concept to class discrimination and the new classification method VPMCD. Distinct class VPM models are designed for each class during the supervised training, using known observations of the system variables. The unknown sample whose class has to be identified is mapped on each of these class VPMs and the attribute values for that sample are reproduced. The sample is classified as belonging to the *class* (nature of variable associations) for which corresponding class VPM gives the best prediction fitness.

As depicted in Fig. 1, for VPMCD algorithm, the supervised learning starts with a given training data set $N\,[n \times p; c]$ with observations (n samples) made on a set of variables (p attributes) to characterize groups of system behavior (c classes). This matrix N is separated into sub matrices for each class of observations, $G_g\,[n_g \times p]$ where n_g is the number of samples in the training set belonging to class g ($g = 1, 2, \ldots, c$), i.e. $\sum n_g = n$. Order r (number of predicting variables in the model) and type as explained in Table 1 are decided by the user. For each feature vector $X_i\,[n_g \times 1]$ ($i = 1, 2, \ldots, p$) in matrix G_g, d sets of remaining variables (feature vectors other than X_i) are selected

Table 1
VPM model types and corresponding details

Model type	Variable predictive model to predict Y (i.e. X_i) (VPM_i) (with j, k and $l \neq i$)	Design matrix (D)	Number of parameters in B (q)
Linear (L)	$b_0 + \sum_{j=1}^{r} b_j X_j$	$[1\ X_1\ X_2 \ldots X_r]$	$1 + r$
Linear + Interaction (LI)	$b_0 + \sum_{j=1}^{r} b_j X_j + \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} b_{kl} X_k X_l$	$[1\ X_1\ X_2 \ldots X_r\ X_1X_2\ X_1X_3 \ldots X_{r-1}X_r]$	$1 + r + {}_rC_2$
Quadratic (Q)	$b_0 + \sum_{j=1}^{r} b_{1j} X_j + \sum_{k=1}^{r} b_{2k} X_k^2$	$[1\ X_1\ X_2 \ldots X_r\ X_1^2\ X_2^2 \ldots X_r^2]$	$1 + 2r$
Quadratic + Interaction (QI)	$b_0 + \sum_{j=1}^{r} b_{1j} X_j + \sum_{k=1}^{r} b_{2k} X_k^2 + \sum_{k=1}^{r-1} \sum_{l=k+1}^{r} b_{kl} X_k X_l$	$[1\ X_1\ X_2 \ldots X_r\ X_1^2\ X_2^2 \ldots X_r^2\ X_1X_2\ X_1X_3 \ldots X_{r-1}X_r]$	$1 + 2r + {}_rC_2$

Note: The number of possible models (Y) for each X_i is $d = (p-1)C_r$ where p is the number of variables and r is predictor order. The model which best predicts X_i is selected as VPM_i.

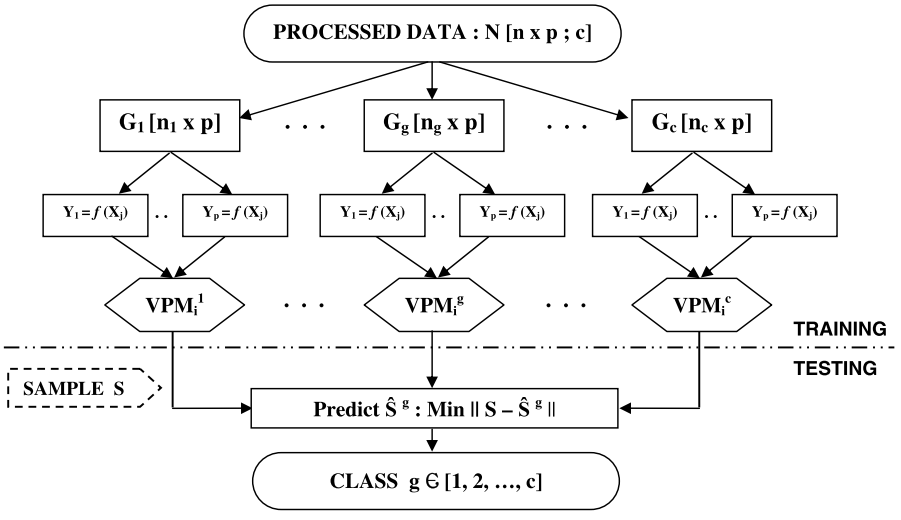

Fig. 1. Schematic flow diagram for VPMCD classification algorithm strategy.

($d = (p-1)C_r$). Depending on the model type and order (r), d different models with design matrices D_j are formulated using equations given in Table 1. The VPMs for each Y_k can be set up as ordinary least squares problem written in general as in Eq. (1)

$$Y_k = D_k \cdot B_k \quad (\text{note that } Y_k = X_i). \tag{1}$$

The set B_k is evaluated using the known feature vector $Y_k = X_i$ and the design matrix D_k obtained from r variables as in Table 1. The ordinary least squares solution is given by Eq. (2)

$$B_k = D_k^{\text{inv}} \cdot Y_k, \tag{2}$$

where D_k^{inv} is the inverse (or pseudo-inverse for non-square D_k) of D_k. The model coefficients B_k are evaluated for all the possible d number of models Y_k for predicting the feature vector X_i. The vector Y_k is then predicted using all the d models to obtain $\hat{Y}_k$. The best model VPM_i is selected as final predictive model for X_i in group g, based on its prediction accuracy given in Eq. (3) as sum of squared prediction errors

$$\text{SSE}_k = \sum (Y_k - \hat{Y}_k)^2; \quad k = 1, 2, \ldots, d. \tag{3}$$

The models and the parameter sets for group g are stored as class models VPM_i^g. Each VPM_i^g ($i = 1, 2, \ldots, p$; $g = 1, 2, \ldots, c$) stores the indices of predictor variables sets (variables used to build models other than X_i) and the model coefficients B_i^g. In the testing scheme, these models are used as class prediction models to classify an unknown sample into its respective structural class. The new sample S ($1 \times p$ features from same N) is projected on the trained models of VPM_i^g. The values in S corresponding to indices of predictor variables are used in VPM_i^g to predict the corresponding X_i value in S. Hence each feature of S is re-predicted to obtain $\hat{S}$ and c different $\hat{S}$ vectors will be obtained. The VPMCD classifier decision is made based on the decision function given in Eq. (4). S is classified as belonging to class g which provides minimum squared prediction error SSE_g.

$$\text{Min}_g \|\text{SSE}\|_g = \left\| \sum_{i=1}^{p} (S_i - \hat{S}_i)^2 \right\|_g; \quad g = 1, 2, \ldots, c. \tag{4}$$

Since all the VPM_i^g ($g = 1, 2, \ldots, c$) are independently trained, they characteristically store the best relations between variables for respective groups. Hence these models themselves directly discriminate the group structures without seeking any decision boundary. The group model VPM_i^g becomes robust as the number of variables increases and can distinguish the groups even if the samples belonging to different groups are closely located in a p dimensional descriptor space. Since the criteria for discrimination are the prediction errors, the VPM based discrimination method does not suffer from the problem of inseparability associated with classifiers based on distance measure or hyper-planes. It should also be noted that once the model type and order r are selected, the VPM parameters can be determined for a given set of training samples. There is no parameter which needs tuning or optimization making the classifier computationally less intense and more robust.

2.3. Data sets

Three widely studied and publicly available data sets are studied to demonstrate the performance of proposed VPMCD algorithm. The binary (2 class: ALL and AML) cancer data set (CANCER), N [72×16063; 2] has been extensively studied by [9–11] as a significant bioinformatics problem. This expression profile data set provides a challenging classification task as the number of attributes (mRNA expression profiles) are very high compared to the number of observations (number of tumor samples). Such problems with $n \ll$ p generally employ a variable selection step before subjecting the samples to classification task. The second data set on HIV [12,13] consists of 208 samples of reverse transcriptase inhibitors (RTIs) belonging to five HIV types (TIBO, HEPT, ITU, DATA and DAPY). The 54 descriptors for each sample are calculated based on the interaction energies of the RTIs with the enzymes. Since the classes TIBO and ITU have statistically insignificant number of samples (2 and 1, respectively) they are not included in the present study. The HIV data set N [205×54; 3] is therefore a classical biological class discrimination problem with large sets of descriptors and enough samples available for training and testing. The third data set [14,15] (protein sub-cellular location, PSL), N [1302×20; 5] consists of protein sequences belonging to five different sub-cellular locations (cytoplasmic, inner membrane, periplasmic, outer membrane and extra cellular). This large data set having non-uniformly distributed samples with 20 amino acid compositions as descriptors is selected because it qualifies as a benchmark multi-class problem. Further details on the organization of these three data sets and descriptions of the features are well documented in [9–15]. These three problems have been analyzed separately in literature using advanced classification techniques like modified LDA [10], CART [12], AdaBoost [13] and support vector machines [9,14]. Hence these data sets present good case studies for comparing the performance of VPMCD with existing superior methods as applied to a variety of computational biology applications.

2.4. VPMCD implementation and testing

The VPMCD algorithm discussed previously has been implemented in MATLAB (version 7.0.4, 2005) [16]. The VPMCD code and the data sets can be made available to readers upon request. All the data sets are preprocessed by mean centering each column and scaling by corresponding standard deviation. The two case studies, HIV and PSL data sets, are directly subjected to VPMCD training whereas a smaller subset of genes are pre-selected before classifying the CANCER data set. The effective sets of genes are selected in three different ways: (i) set of 18 genes as given in [11], (ii) based on Fisher criteria (FC) defined as the ratio of inter-class to the intra-class variations in the gene expression levels [2,3,17] and (iii) a set of 4 genes based on SVM selection [18]. The new classifier performance is validated using re-substitution (self-consistency), cross validation (leave one out – LOOCV and multifold – nFCV) and new sample prediction tests [2]. These tests effectively bring out various objectives of data classification and indicate the stability, robustness and generality of the discriminating method. Different classification methods, analyzed based on the same classifier testing approaches discussed above, are used to compare the performance of the new algorithm wherever applicable for each data set. The results for the comparison methods are directly reported from the relevant literature (ULDA [10], CART [12], AdaBoost [13], SVM [9,14,18] and k-TSP [11]). Readers are advised to consult these references for the settings selected and optimization of parameters used for the respective methods.

3. Results and discussion

Table 2 highlights different test results for the three data sets considered in this study. All the four model types and different predictor order r are used to perform the initial resubstitution tests. The VPM type and order r giving highest self consistency accuracy for a specific data set are selected. These settings are mentioned in the first column of Table 2. The classification accuracies are indicated as the overall percentage of samples predicted correctly during testing. For cross validation test, the results are mean accuracies over several random iterations as indicated in the first column.

Results for resubstitution test clearly indicate the excellent learning ability of the new classifier across different sizes and types of data sets selected. This establishes the notion that interaction amongst the variables (gene interactions for CANCER, activity relations for HIV and amino acid interactions in PSL) can effectively capture the characteristic of different classes and can be used to distinguish them. Linear VPMs with $r = 1$ provide best discrimination of tumor classes (AML/ALL). This indicates one to one dependency of genes in establishing the tumor characteristics. 100% performance for CANCER data set with only 18 out of 16063 genes [11], 7 FC genes (accession ids: U46499, M27981, M23197, M84526, X17042, X95735 and L09209) and 4 SVM selected genes (accession ids: M27891, M19507, L20688 and Y00787_s) [18] confirms the compatibility of VPMCD to different variable selection methods. The cross validation and new sample test results provide better insights to the superiority of the new proposed

Table 2
Training and validation test results for VPMCD on three data sets in comparison with existing methods from literature

Dataset/test/(settings)	VPMCD % accuracy	Comparison with best methods[a]
CANCER: (*Linear VPM with* $r = 1$)		
Resubstitution (18 genes [11]/7 genes – FC)	100	–
5FCV (18 genes [11], mean of 25 iterations)	100	–
3FCV (18 genes [11], mean of 50 iterations)	99.45	97.67 (ULDA [10])
LOOCV (18 genes [11]; 11 genes FC)	100; 98.61	95.83 (k-TSP [11]) 97.22 (enSVM [9])
Sample test (train/test: 38/34, 4 genes [18])	100	97 (SVM [18])
HIV: (*Linear VPM with* $r = 2$)		
Resubstitution	100	91 (CART [12])
10FCV (mean of 10 iterations)	95.79	74 (CART [12])
Sample test (train: 128, test: 80 samples)	96.25	90 (Adaboost [13]) 83.7 (CART [13])
PSL: (*QI VPM with* $r = 5$)		
Resubstitution	90.9	–
5FCV (mean of 10 iterations)	78.81	78.6 (SubLoc SVM [14])

[a]Best method as suggested by the respective references for corresponding dataset and for the same performance test as reported in first column of respective row.

method. Compared to the best available SVM classifier, VPMCD method efficiently predicts the classes for untrained test samples. This indicates the stability of model based approach compared to the optimum hyper-plane based SVM method. *k-TSP* [11] based gene selection and tumor classification approach is reported as the best method overall. The improved and complete classification result with the same set of genes as selected by k-TSP establishes the advantages of the new VPMCD classifier especially for $n \ll p$ problems. It is also observed during the 3FCV tests that VPMCD has only ±1.3% standard deviation in prediction accuracies over 50 iterations as against 2.4% reported by [10] using uncorrelated linear discriminant analysis (ULDA) method. This indicates the robustness of VPMCD mainly due to the gene interactions considered in the proposed algorithm. The sets of linear, univariate ($r = 1$) VPMs selected for CANCER data set make the VPMCD algorithm faster as compared to the kernel based SVM method [9].

Full variable set of 54 descriptors with 208 samples are used to analyze the HIV data. Linear bivariate models give the best prediction accuracies overall. This shows the importance of multivariate enzyme interactions in deciding the activity of inhibitors. VPMCD provides a significant improvement in all the test results compared to the successfully employed decision tree methods. The effectively designed activity interaction models over a large set of descriptors can distinguish the HIV classes better than univariate comparison based CART methods. New independent test set with 80 samples (unused during training) is predicted with additional 6% accuracy compared to Adaboost and 13% more accuracy compared to the normal CART method. The VPMCD algorithm is simpler in construction and faster in training and testing as compared to iterative Adaboost CART approach. These observations support the ability of VPMCD method to efficiently learn and detect various classes of high dimensional data.

A detailed analysis on PSL data reveals that quadratic interaction (QI) models with variable predictor order of 5 can effectively separate the cellular locations of large sets of proteins. This reveals the biological significance of interacting amino acids in long peptide chains in deciding their structure, cellular location and in turn their functions. The effect of model type and order (r) on the VPMCD performance is presented in Fig. 2a and b, respectively. The increase in accuracy with increasing r value highlights the multiple nonlinear interactions among the amino acid molecules which characterize the mobility and binding properties of proteins. The 5FCV test result is comparable to the best available PSL classifier (SubLoc [14,19]) employing SVM method on only the amino acid composition data. The unique advantage of VPMCD approach is its ability to addresses the multi-class problem in single training step unlike SVM method which employs $c*(c-1)/2$ binary classifiers for a 'c' group problem. The classification performance for a large set of multi-class data can be improved with additional descriptors which are known to influence the protein characteristics [14,15,19]. This is evident from the trial results (not shown) obtained with 400 dipeptide compositions $N[1302 \times 420; 5]$ as additional descriptors. The resubstitution result for VPMCD (QI; $r = 1$) significantly increases from 79% (Fig. 2b) to 90% and for 5FCV (QI; $r = 1$) the classification accuracy increases from 74% (Fig. 2b) to 77%. This indicates possible improvement results for multivariate PSL data with increasing descriptor space.

These case studies support the proposed new VPMCD method as a strong and potential tool for variety of classification applications in computational biology. During the analyses, identical model type and order (r) is selected for all the classes in training step. The study can be extended analyzing the effect of selecting separate model types and orders for each class. Also, models used here are limited only to four basic types. The results can be improved with more accurate models defined using a priori system knowledge or by new modeling techniques like genetic programming [20]. The VPMs are optimized based on their re-predictive capabilities during training. Alternatively, the best VPM_i can be selected based on untrained sample validation accuracy during training which can further improve the performance of VPMCD for nFCV and LOOCV tests. Factors like availability of continuous and dependent features, sufficient number of observations in each group to provide higher statistical strength to parameters estimated during the training algorithm can affect the performance of VPMCD. Further investigations on these aspects can equip the new variable interaction model based approach with capa-

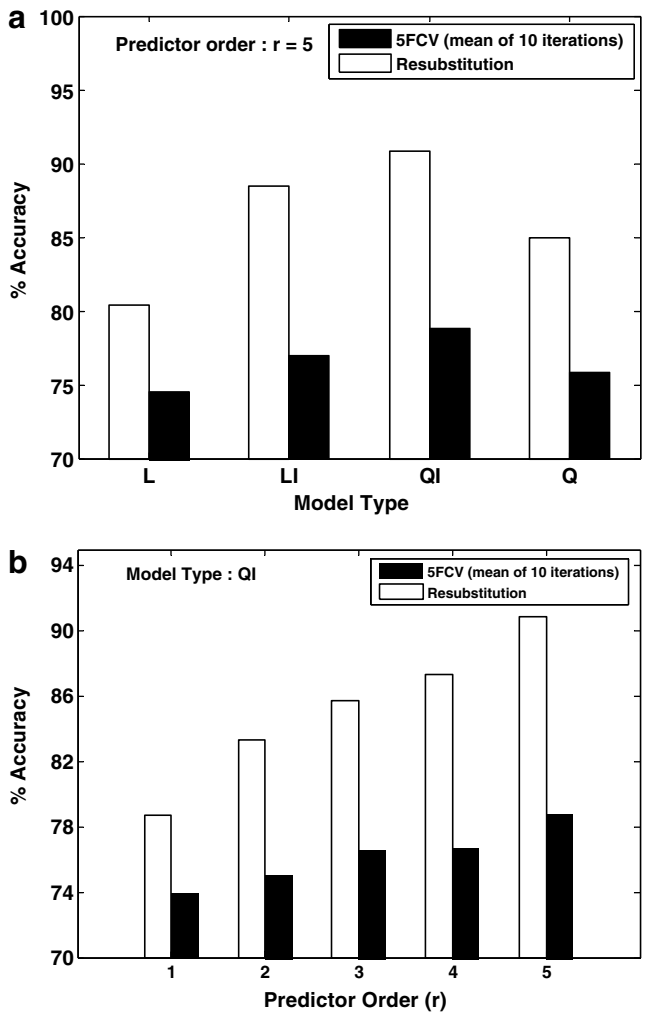

Fig. 2. Analysis of resubstitution and 5FCV results for PSL data. (a) Effect of different model types on accuracy for predictor order $r = 5$. (b) Effect of different values of r on accuracy using QI type VPM models.

bilities to quickly and effectively solve large scale classification problems with relatively less computational effort.

4. Conclusions

A new class discriminant method VPMCD based on variable interaction models is proposed. The performance of the new classifier is analyzed using three well studied data sets that have a range of biological and statistical significance. The results obtained based on different classification tests establish the overall superiority of VPMCD method compared to several existing methods reported to perform the best for respective problems. The performances for more rigorous cross validation tests reveal the stability and robustness of the new method. Investigations on the effect of predictor order and model type establish the inherent strength of the new method in capturing the discriminative variable interactions. The performance of this new approach can be enhanced with improvements suggested in Section 3. With detailed investigation on the effect of sample size and class distribution, the new method can be extended to more complex classification problems involving biological systems.

References

[1] Sokal, R.R. and Rohlf, F.J. (1994) Biometry: The Principles and Practice of Statistics in Biological Research, 3rd ed, Freeman, New York.
[2] Duda, R.O., Hart, P.E. and Stork, D.G. (2000) Pattern Classification, Wiley, New York.
[3] McLachlan, G.J. (1992) Discriminant Analysis and Statistical Pattern Recognition, Wiley–Interscience, New York.
[4] Vapnik, V. (1998) Statistical Learning Theory, Wiley–Interscience, New York.
[5] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) Classification and Regression Trees, Wadsworth International Group, Monterey, CA.
[6] Kurgan, L.A. and Homaeian, L. (2006) Prediction of structural classes for protein sequences and domains – Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recog. 39, 2323–2343.
[7] Pearl, J. (2000) Causality: Models, Reasoning, and Inference, Cambridge University Press, Cambridge, UK.
[8] Beck, J.V. and Arnold, K.J. (1977) Parameter Estimation in Engineering and Science, Wiley, New York.
[9] Peng, Y. (2006) A novel ensemble machine learning for robust microarray data. Comput. Biol. Med. 36, 553–573.
[10] Ye, J., Li, T., Xiong, T. and Janardan, R. (2004) Using uncorrelated discriminant analysis for tissue classification with gene expression data. IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (4), 181–190.
[11] Tan, A.C., Naiman, D.Q., Xu, L., Winslow, R.L. and Geman, D. (2005) Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics 21 (20), 3896–3904.
[12] Daszykowski, M., Walczak, B., Xu, Q.S., Daeyaert, F., de Jonge, M.R., Heeres, J., Koymans, L.M.H., Lewi, P.J., Vinkers, H.M., Janssen, P.A. and Massart, D.L. (2004) Classification and regression trees: studies of HIV reverse transcriptase inhibitors. J. Chem. Inf. Comput. Sci. 44, 716–726.
[13] Zhang, M.H., Xu, Q.S., Daeyaert, F., Lewi, P.J. and Massart, D.L. (2005) Application of boosting to classification problems in chemometrics. Anal. Chim. Acta 544, 167–176.
[14] Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusna'dy, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K. and Brinkman, F.S.L. (2003) PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res. 31 (13), 3613–3617.
[15] Su, C.Y., Lo, A., Chiu, H.S., Sung, T.Y. and Hsu, W.L. (2006) Protein subcellular localization prediction based on compartment specific biological features, in: Proceedings of IEEE CSB2006 Computational Systems Bioinformatics Conference.
[16] MATLAB 7.0.4 Release 14 (2005) The MathWorks Inc., Natick, MA.
[17] Guyon, I., Weston, J. and Barnhill, S. (2002) Gene selection for cancer classification using support vector machines. Mach. Learn. 46, 389–422.
[18] Fua, L.M. and Fu-Liu, C.S. (2004) Multi-class cancer subtype classification based on gene expression signatures with reliability analysis. FEBS Lett. 561, 186–190.
[19] Hua, S. and Sun, Z. (2001) Support vector machine approach for protein sub cellular localization prediction. Bioinformatics 17, 721–728.
[20] Koza, J.R. (1992) Genetic Programming: On the Programming of Computers by means of Natural Selection, MIT Press, Cambridge, MA.